Detect the Churn

Churn Reduction Analysis in R and Python

Lakshveer Singh

5/30/18

Table of Contents

1 Executive Summary

Key Findings

- Dived into the groups of data based on their total usage of minutes and calls and total charge which gave a good insight about the behavior of the churned customers.
- Provided the insights with the graphs in the exploratory analysis.
- Used SMOTE to oversample the minority data which is churned customers.
- Analysis includes 2 model. I have tried SVM Classifier and Logistic regression to train the data. The best result is given by SVM Classfier.
- The Sensitivity of the model is 81.25 and Accuracy is 77.98. We are concerned about sensitivity i.e. no of true positive cases correctly classified (Churned out customers) for our problem which looks good enough.

2 Introduction

2.1 Background

Churn (loss of customers to competition) is a problem for companies because it is more expensive to acquire a new customer than to keep your existing one from leaving. In this implementation the focus is on both predicting as accurate as possible whether a person is going to churn and on determining important factors that influence churners.

2.2 Problem Statement

The objective of this Case is to predict customer behaviour. We are providing you a public dataset that has customer usage pattern and if the customer has moved or not. We expect you to develop an algorithm to predict the churn score based on usage pattern.

The predictors provided are as follows:

● account length
● international plan
● voicemail plan
● number of voicemail messages
● total day minutes used
● day calls made
● total day charge
● total evening minutes
● total evening calls
● total evening charge
● total night minutes
● total night calls
● total night charge
● total international minutes used
● total international calls made
● total international charge
● number of customer service calls made

Target Variable : move: if the customer has moved (1=yes; 0 = no)

3 Exploring Data

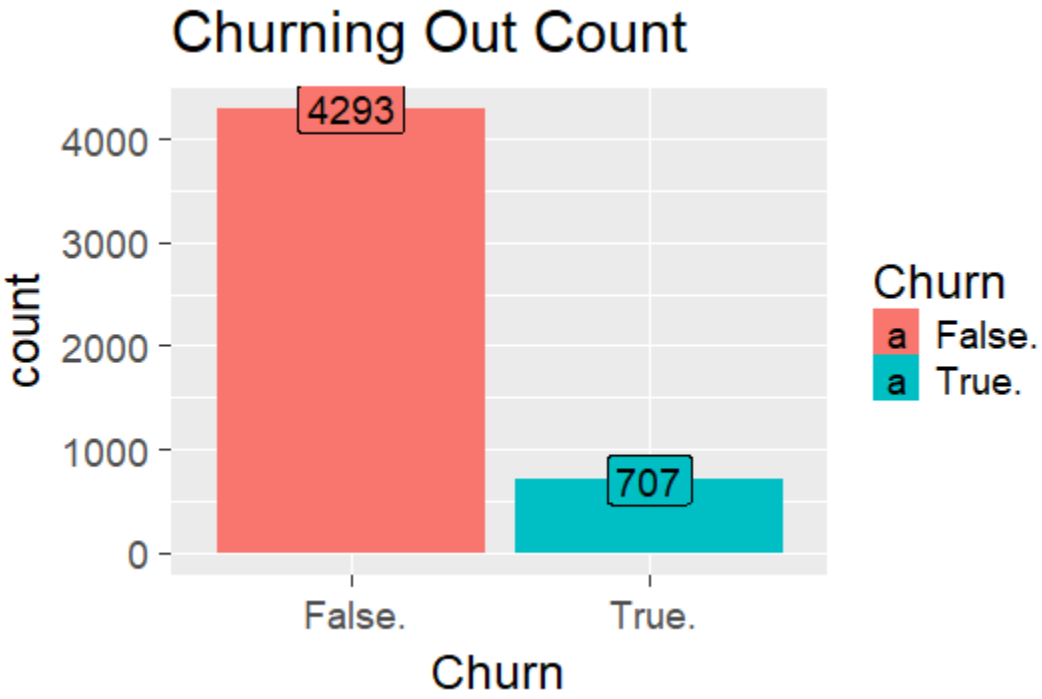
3.1 - 3.2 Data size and structure

```
'data.frame':                   5000 obs. of  21 variables:
$ state                         : chr  "KS" "OH" "NJ" "OH" ...
$ account.length                : int  128 107 137 84 75 118 121 147 117 141 ...
$ area.code                     : int  415 415 415 408 415 510 510 415 408 415 ...
$ phone.number                  : chr  " 382-4657" " 371-7191" " 358-1921" " 375-9999" ...
$ international.plan            : chr  " no" " no" " no" " yes" ...
$ voice.mail.plan               : chr  " yes" " yes" " no" " no" ...
$ number.vmail.messages         : int  25 26 0 0 0 0 24 0 0 37 ...
$ total.day.minutes             : num  265 162 243 299 167 ...
$ total.day.calls               : int  110 123 114 71 113 98 88 79 97 84 ...
$ total.day.charge              : num  45.1 27.5 41.4 50.9 28.3 ...
$ total.eve.minutes             : num  197.4 195.5 121.2 61.9 148.3 ...
$ total.eve.calls               : int  99 103 110 88 122 101 108 94 80 111 ...
$ total.eve.charge              : num  16.78 16.62 10.3 5.26 12.61 ...
$ total.night.minutes           : num  245 254 163 197 187 ...
$ total.night.calls             : int  91 103 104 89 121 118 118 96 90 97 ...
$ total.night.charge            : num  11.01 11.45 7.32 8.86 8.41 ...
$ total.intl.minutes            : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
$ total.intl.calls              : int  3 3 5 7 3 6 7 6 4 5 ...
$ total.intl.charge             : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...
$ number.customer.service.calls: int  1 1 0 2 3 0 3 0 1 0 ..
$  Churn                        : chr  " False." " False." " False." " False." ...
```

Division of train and test data:

```
: Train   : Test
Observations: 3333   : 1667
Percentage :   67       : 33
```

3.3 Class Imbalance



There are 707 customers who churn out and 4293 customers retained.

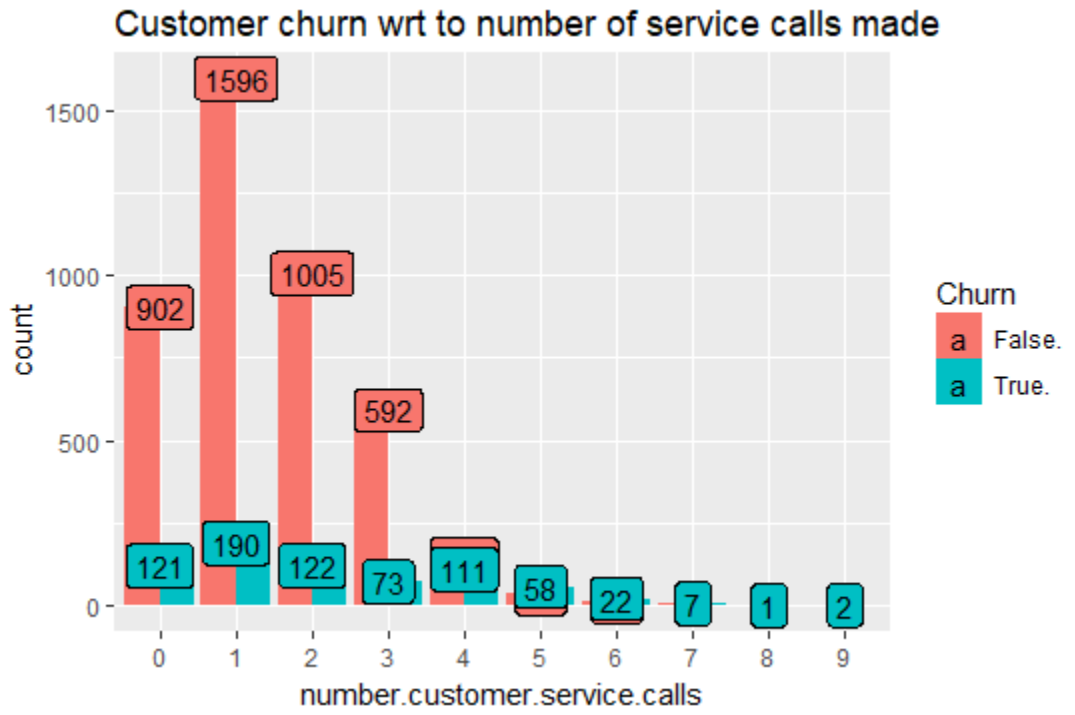
Customer Churn Percentage	14.14
Customer Retained Percentage	85.86

3.4 Completeness of data:

	Missing_Value_Count
state	0
account.length	0
area.code	0
phone.number	0
international.plan	0
voice.mail.plan	0
number.vmail.messages	0
total.day.minutes	0
total.day.calls	0
total.day.charge	0
total.eve.minutes	0
total.eve.calls	0
total.eve.charge	0
total.night.minutes	0
total.night.calls	0
total.night.charge	0
total.intl.minutes	0
total.intl.calls	0
total.intl.charge	0
number.customer.service.calls	0
Churn	0

Checking the distribution of values for the calling related features.

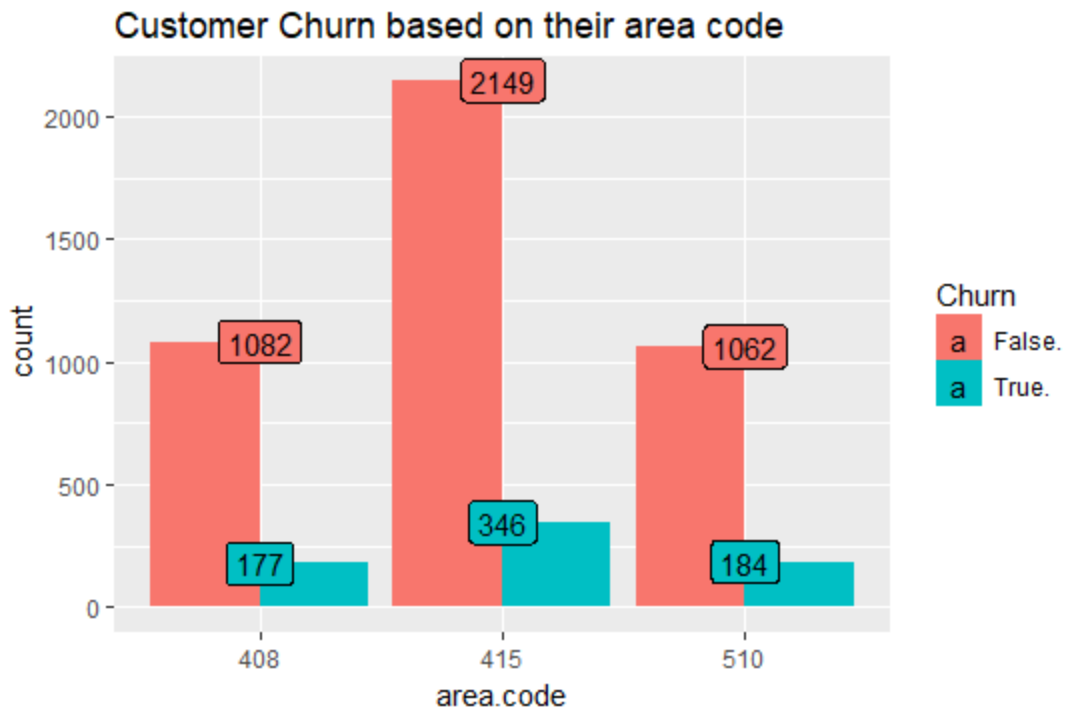
3.5 Exploring some of the most important variables



Customer service calls is an important variable as customers make calls only when they are facing some issues. As we can see in the above plot the max customers who churn out seems to be facing some issue/concern which made them call customer care for resolution. This adds to the customer churn rate as it tells us that customers were not happy/satisfied with the service.

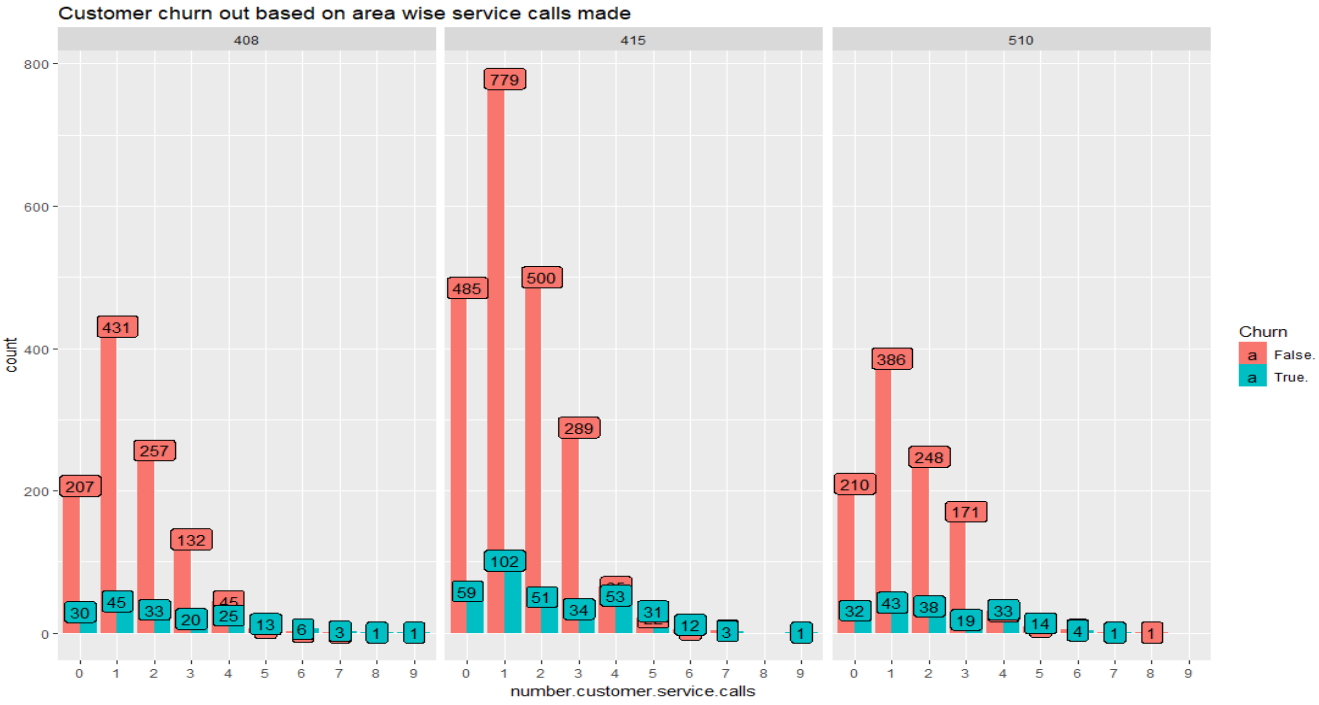
The number of customer who churn out who made 1 or more calls are 586 which is almost 83 percent of customer from the churned out list.

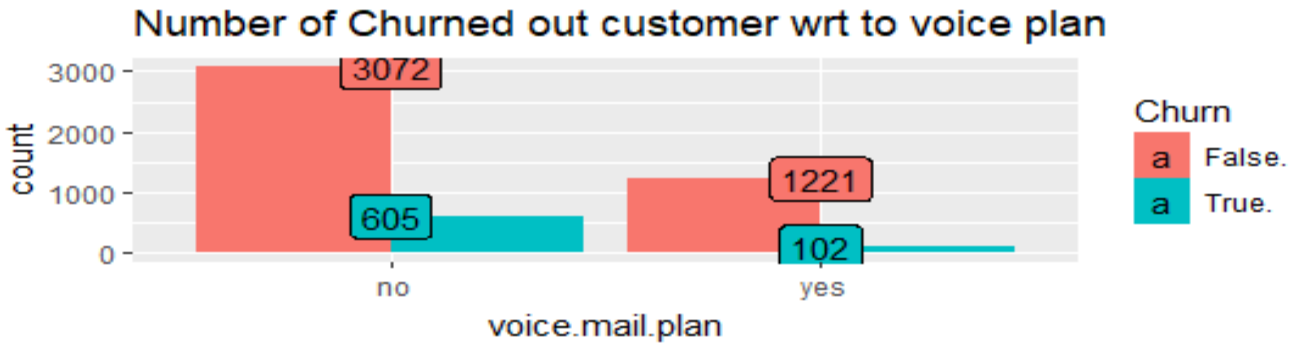
Looking at the above plot we can say that customer service calls have a good impact in customer churning out.

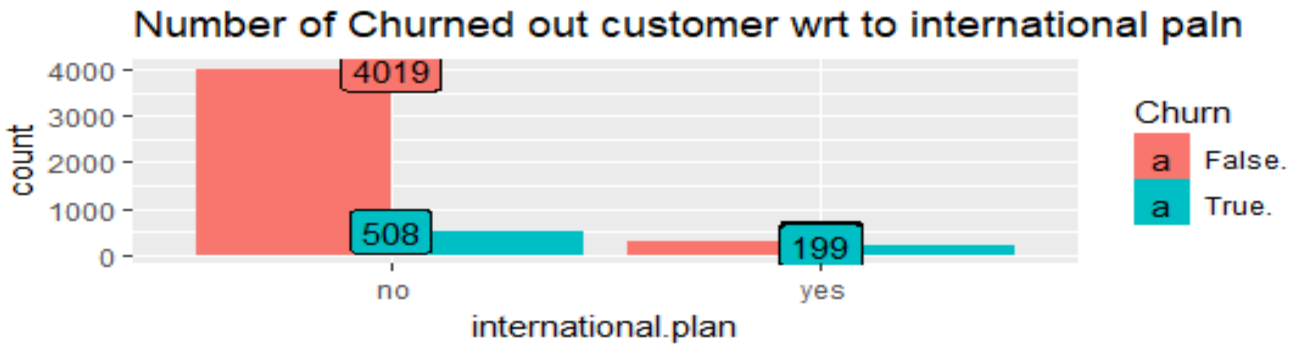


The plot gives us the churn out customers based on their living area. As we can see out 50% of the customers reside in area 415 and half the customers from the customers who churned are from this area. Looking at this data we can also say that the service problem people are having is not area specific like week network/connection/disconnection but seems to be a generic problem.

Let us dig a bit more down the customers with area code and service calls made.

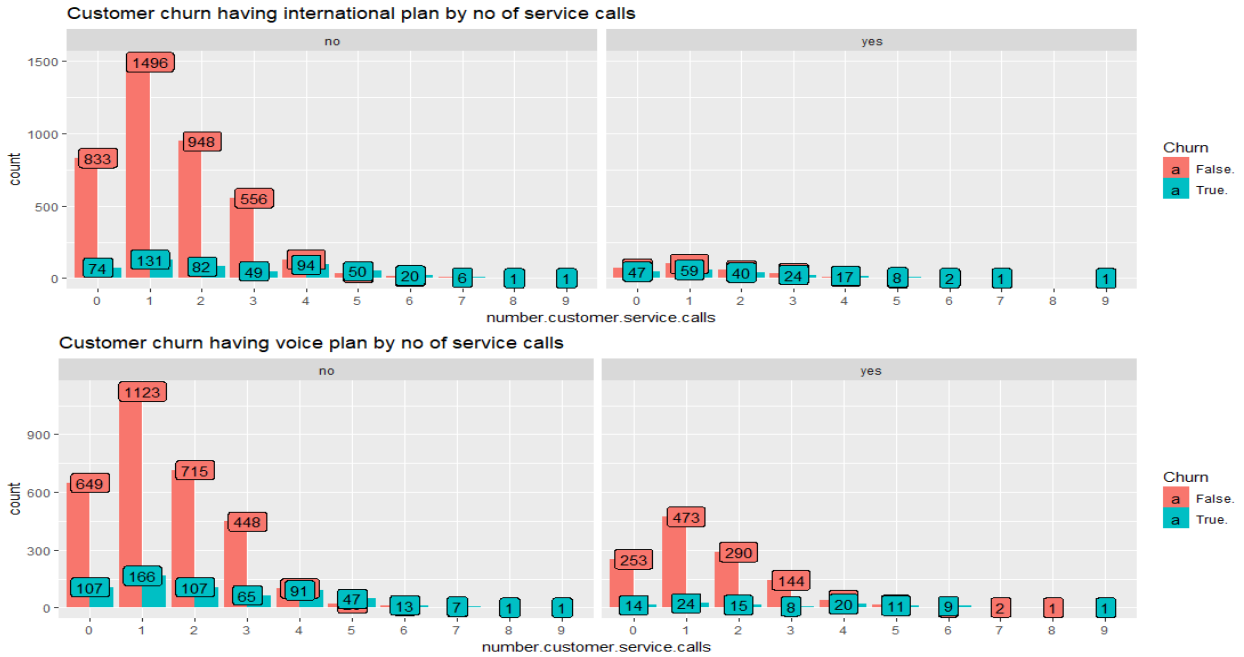






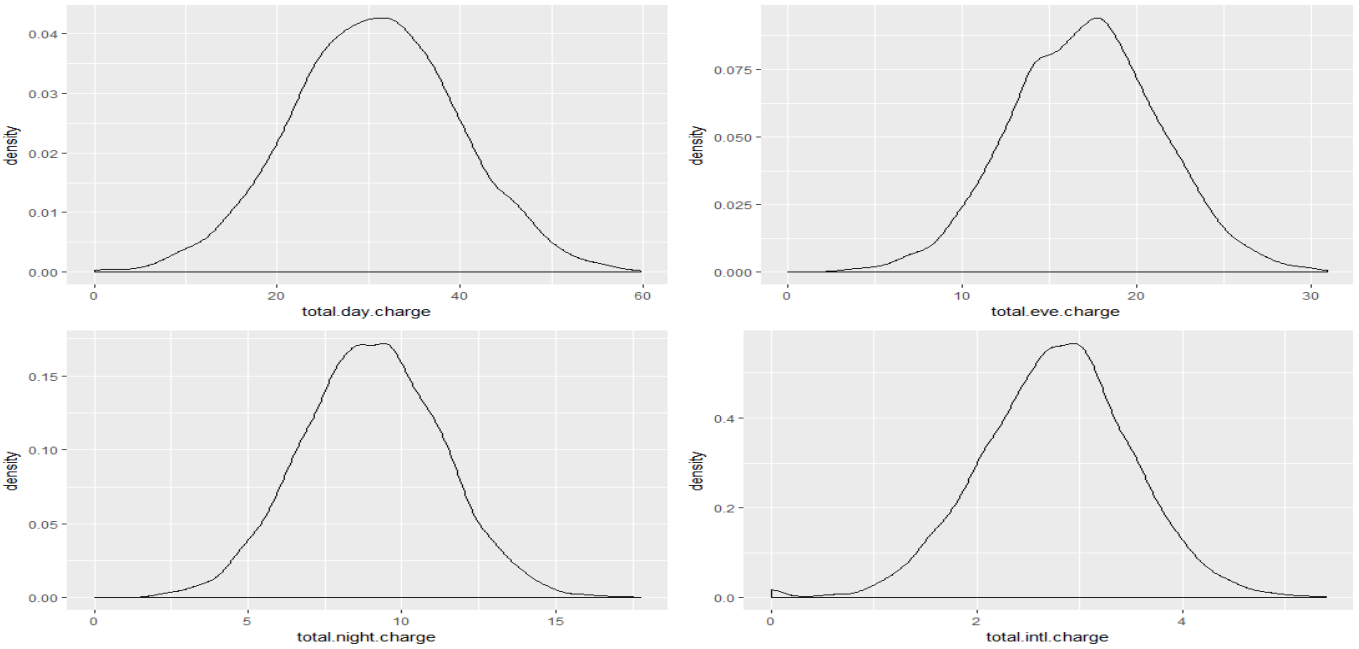
We can see that the customer who were having international plan i.e 463 customer out of those 199 churn out.

Let us analyze international and voice plan with customer service calls to see if we find something interesting.



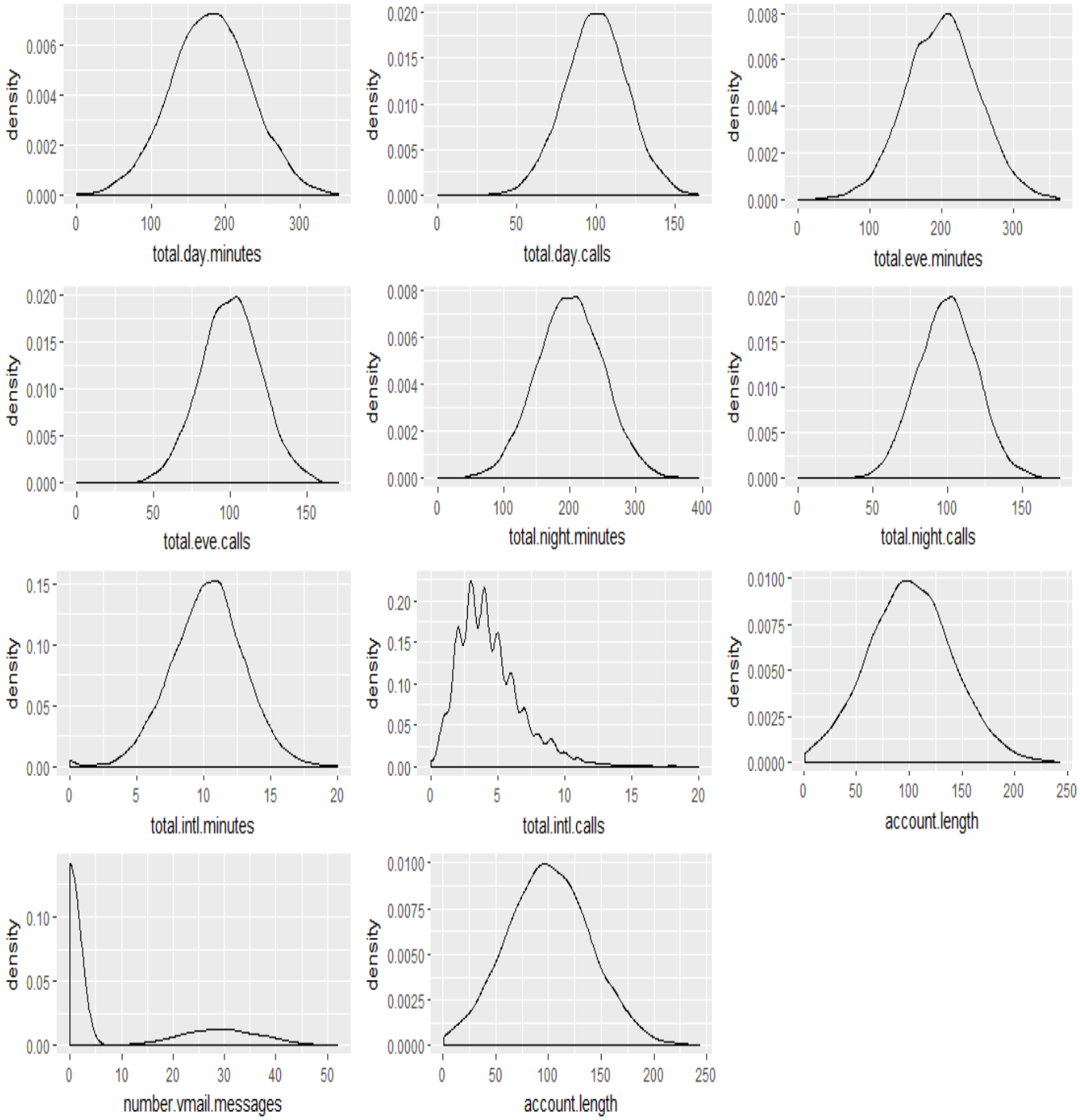
3.6 Density Plot: Distribution of data points

Checking the distribution of values for the calling related features. The charges dataset seems to be normally distributed.



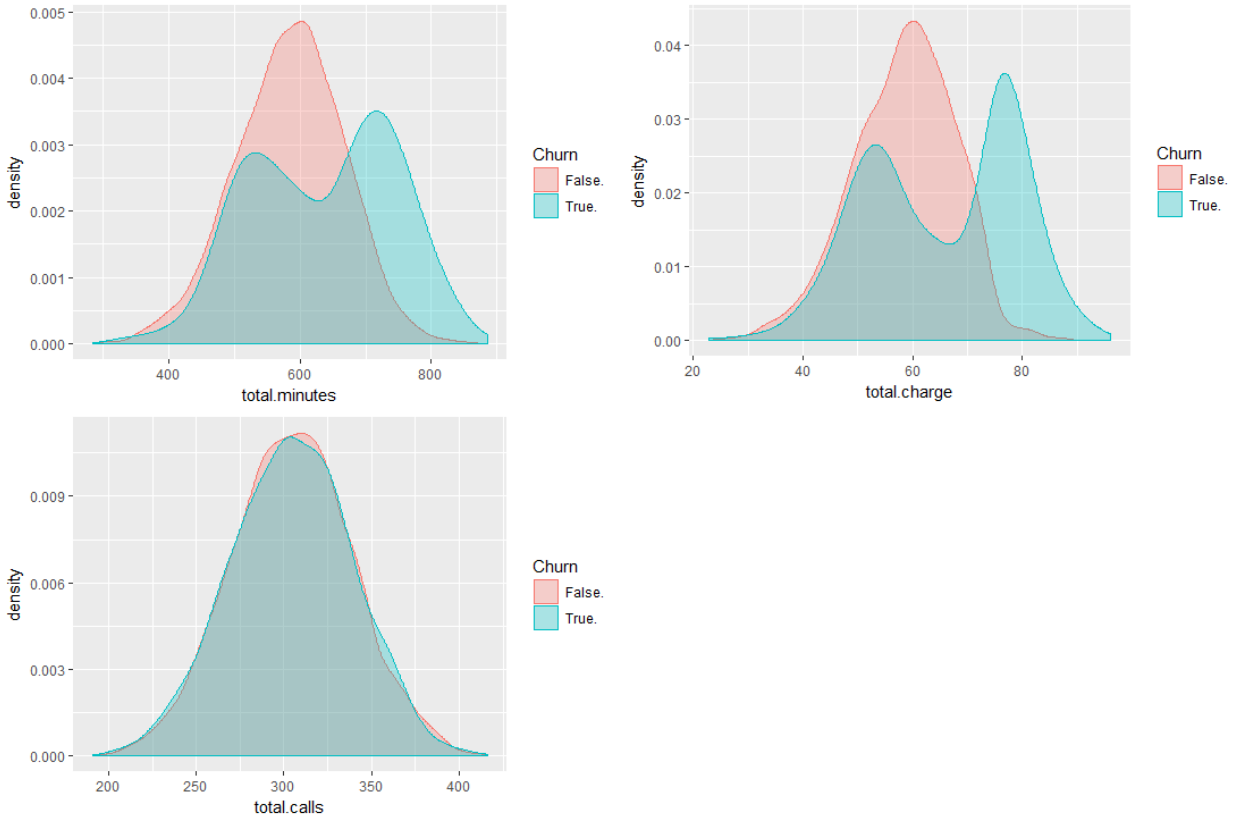
The plot shown the normal distribution for all churn and retained customer based on day charge expect the day charge.

The below plot shown the minutes and calls expect for customer service calls, number of voice mail messages and international calls.



Feature Engineering Variable distribution:

- Total Minutes
- Total Calls
- Total Charges
- Prefix



Insights from the plot:

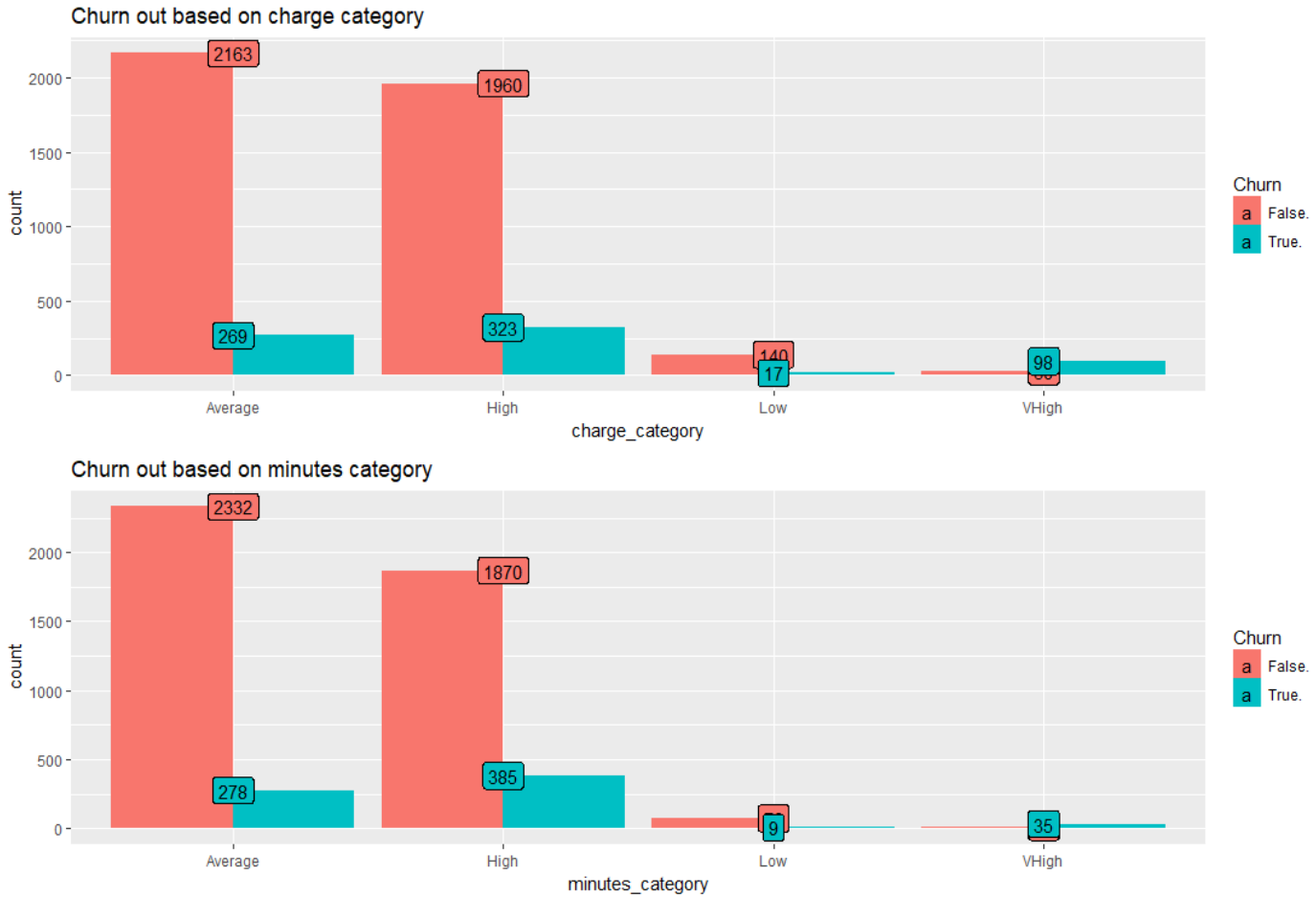
- We can see that the customers who didn't churn out follows a normal distribution for total minutes, charges and calls.
- The plot shows that the total minutes and total charge for the customer who churn out follows a bimodal distribution. Total calls has normal distribution.
- The bimodal nature of total minute and charges shown 2 different groups of customers who churn out means there are 2 modes i.e. most frequent numbers.
- The local maximum for total minutes is around 520 and global maximum around 720. The local maximum for total charge is around 53 and global maximum is around 77.

4 Feature Engineering: Explore More

4.1 Checking for Indirect Features

4.2 Exploring engineered features with response variable

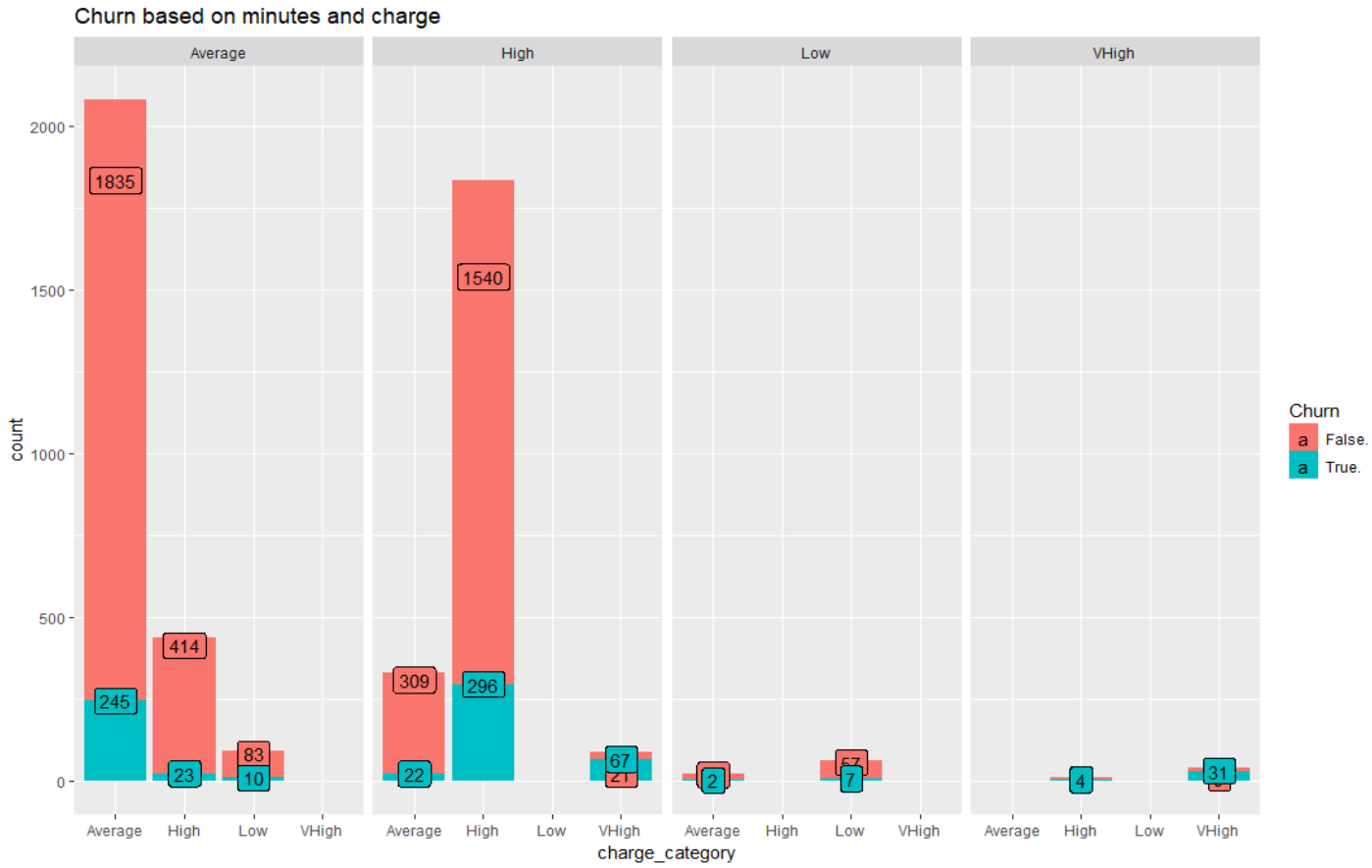
Let's create a categorical variable for total charge and minute and analyze them.



Insights:

- In charge category, the customer churn for high and very high is 59.54 % of the total churn out customers.
- 94.3 % customers fall under Average and High charge category.
- 83.74 percent from the total churn out customers falls under Average and High charge category.
- In minute's category, the customer churn for high and very high is 59.40% of the total churn out customers.
- 97.3% of the customers falls under Average and High minute's category.
- 93.7% from the total churn out customers falls under Average and High charge category.

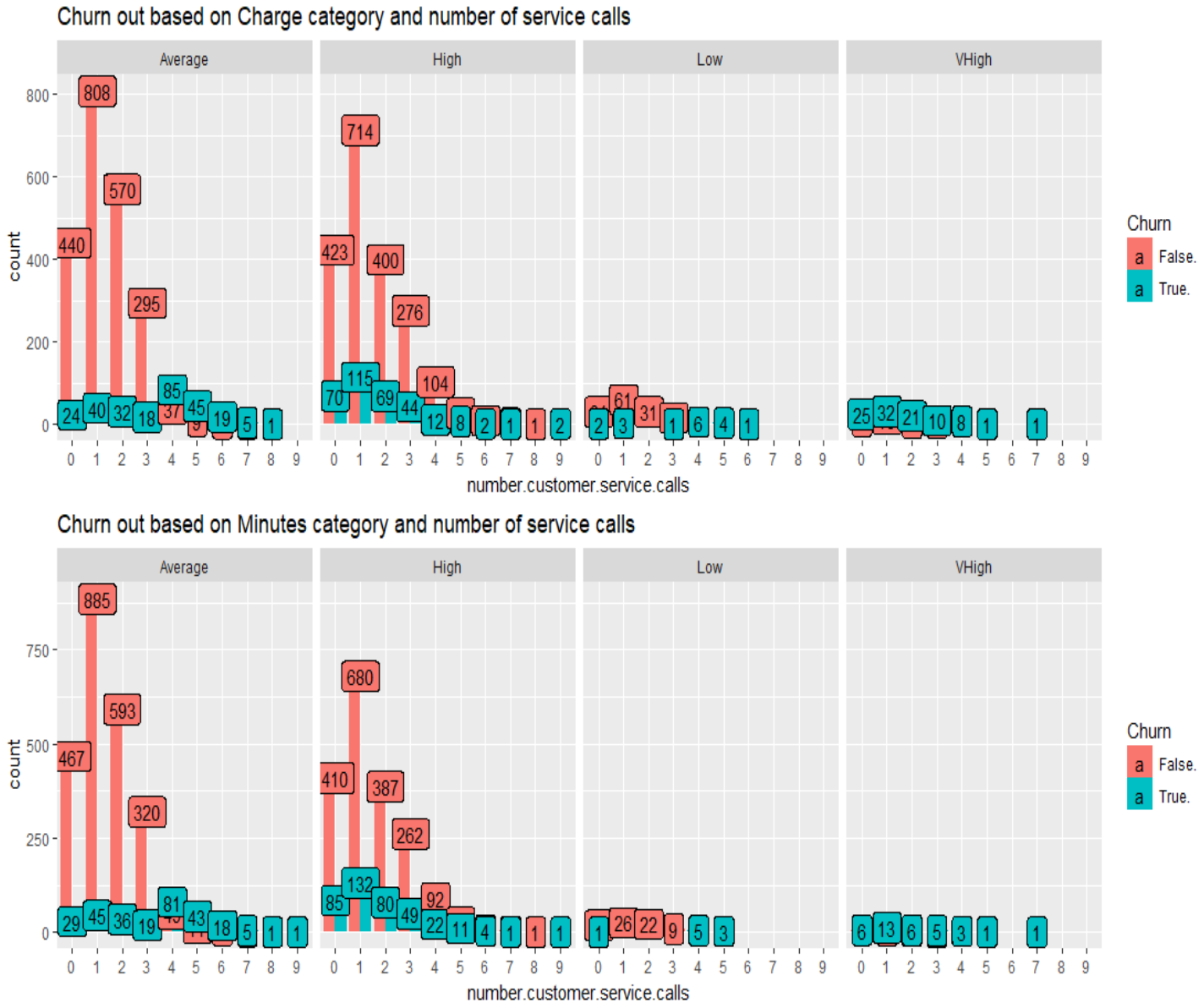
Do the customers with less minutes paying high?



Insights:

- There don't seem to be a problem where people with less minutes are paying high. Max customer in average minutes category falls under average charge category which seems reasonable. The once being charged high for less average minutes usage may have international minutes/plans.

Do the customers paying average and high are suffering with customer network issue?



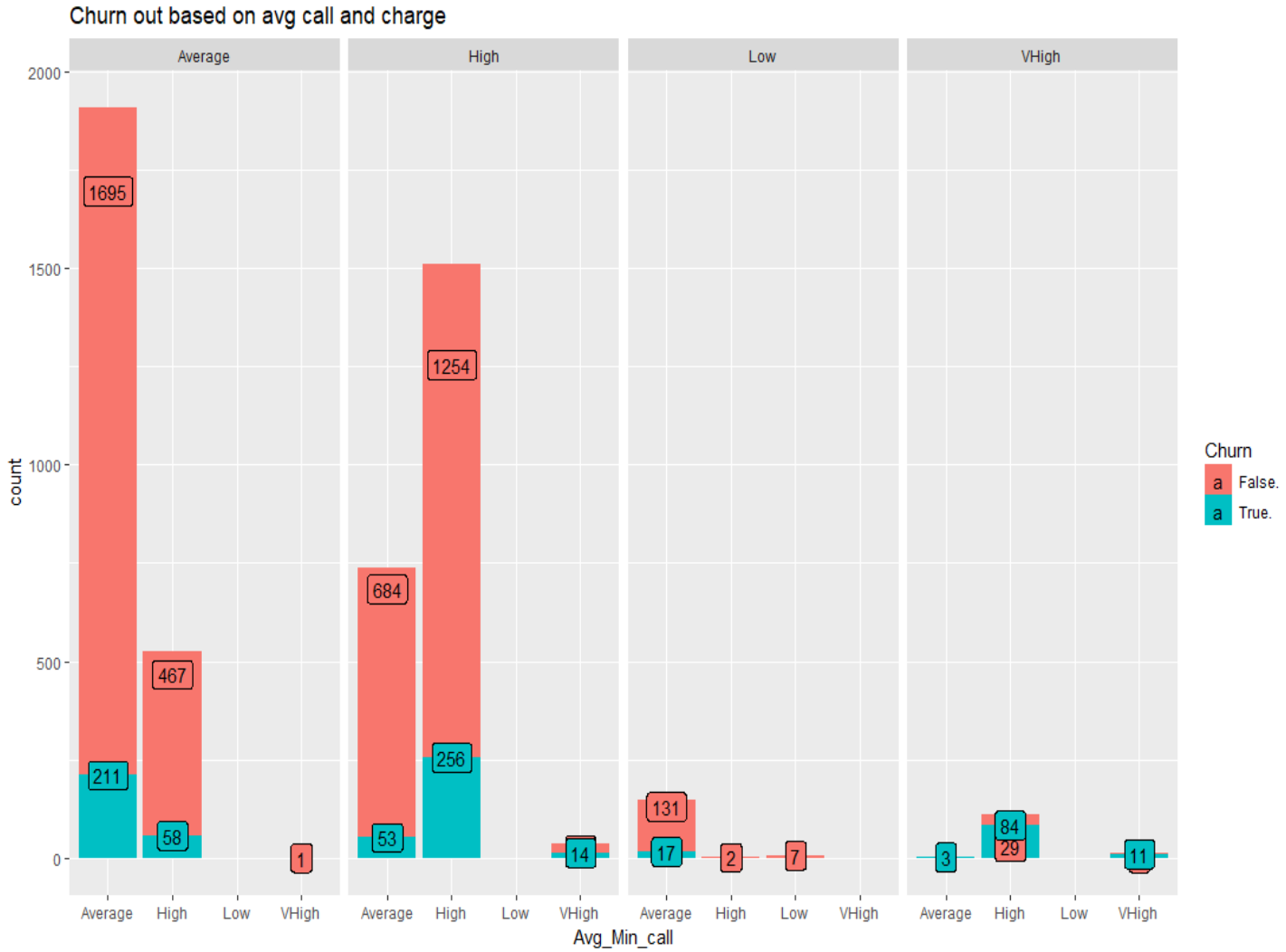
Insights:

- 46.11% from the total churn out customer who falls under high and very high charge category has made 1 or more customer service calls. Looks like they were upset with the service or the resolutions of their issue.
- 54.31% from the total churn out customer who falls under very high, high and average category has made 2 or more customer service calls.

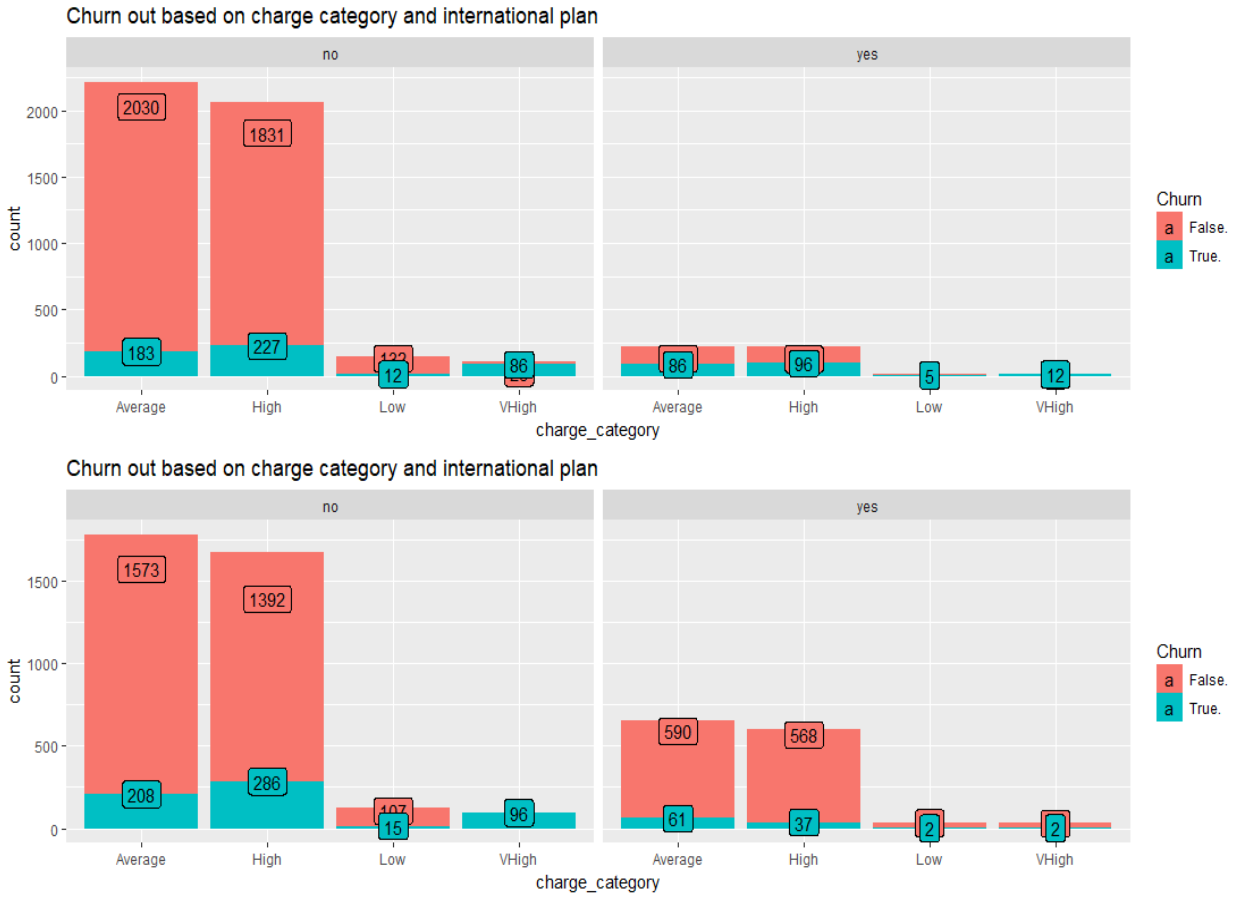
Summary of total calls, total minutes and total charge

```
Total Calls:

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
191.0   282.0   305.0   304.6   328.0   416.0

Total Charge
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
22.93   52.48   59.51   59.49   66.39      96.15


Total Minutes
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
284.3   531.2   592.4   591.6   652.0      885.0
```

Does people with less average min/call paying high?

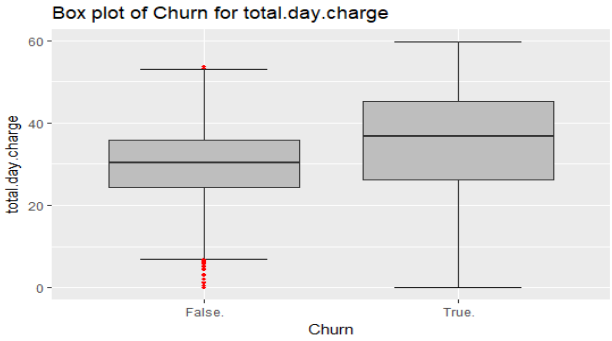


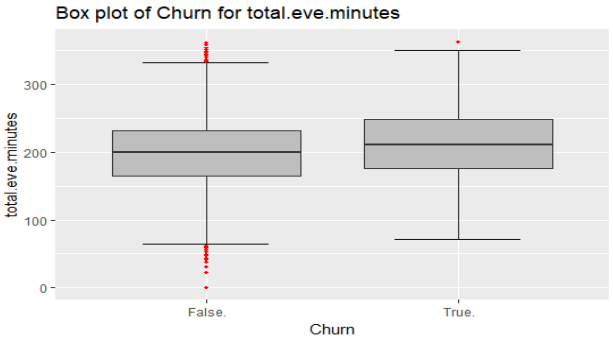
Does people with high charge have international plan?

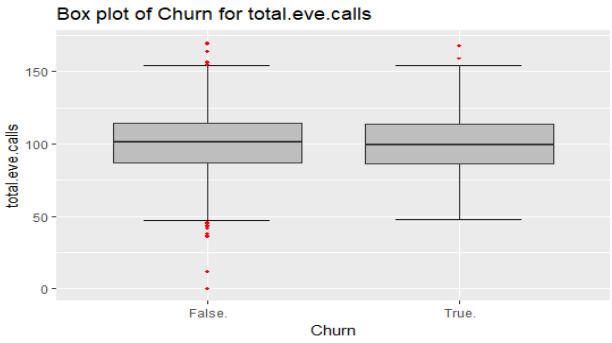


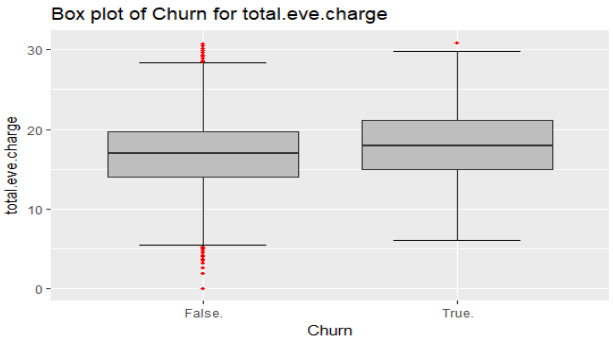
5 Data Pre-processing

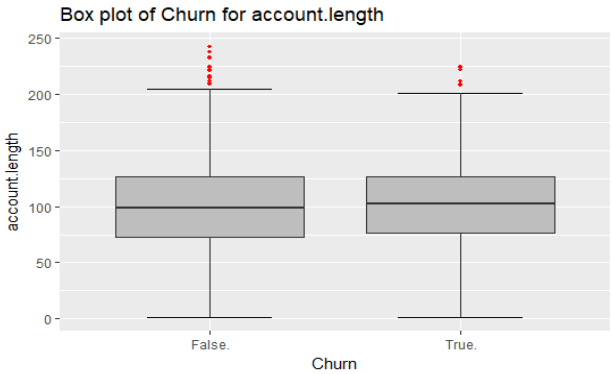
5.1 Anomaly Detection

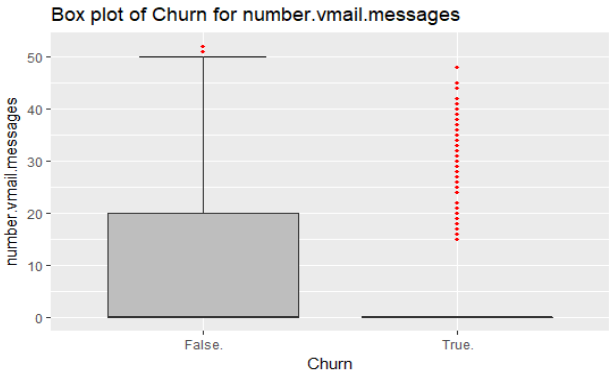


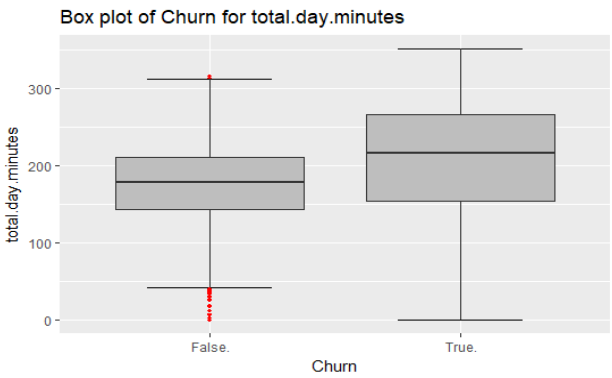


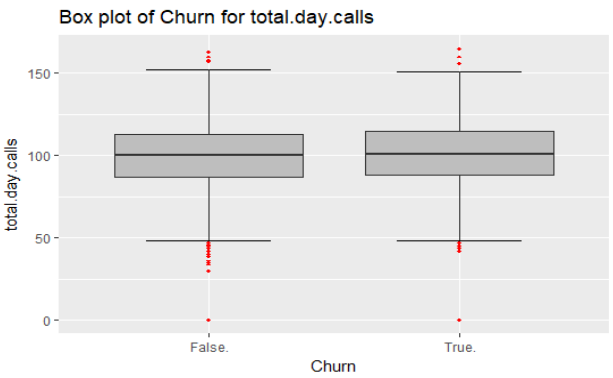


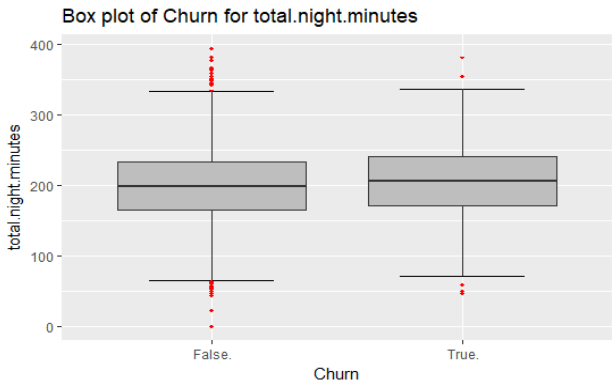


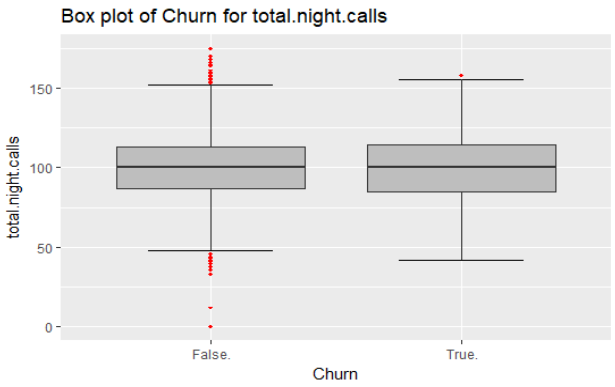


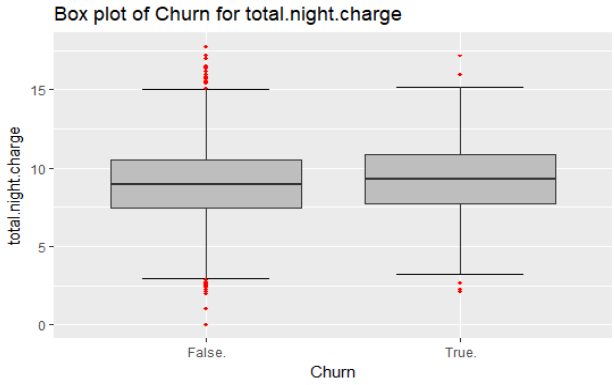


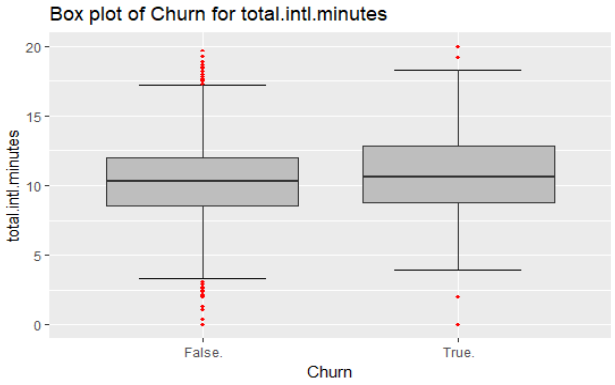


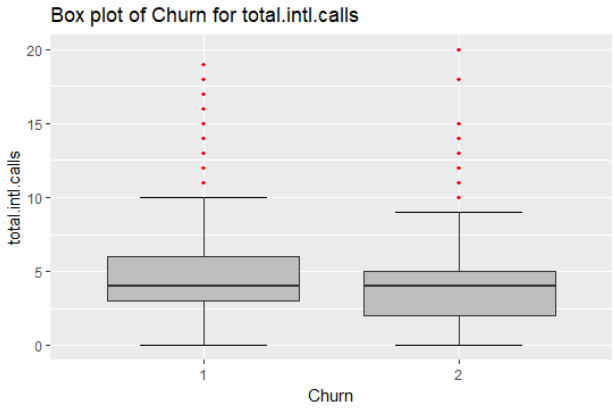


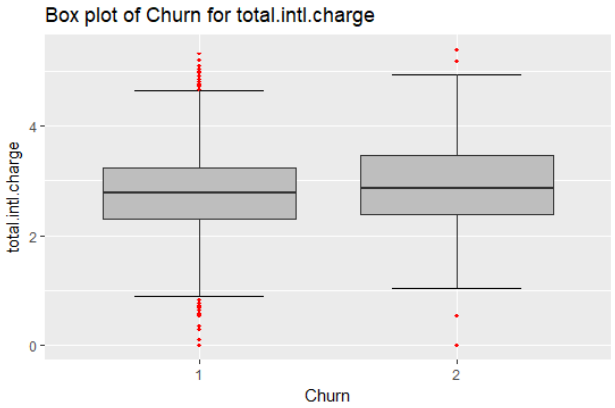

Box plot of Churn for total.night.minutes
total.night.minutes
400
300
200
100
0
False.
True.
Churn

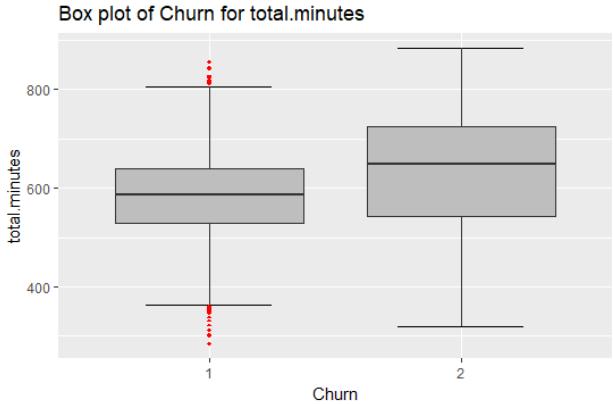

Box plot of Churn for total.night.calls
total.night.calls
150
100
50
0
False.
True.
Churn

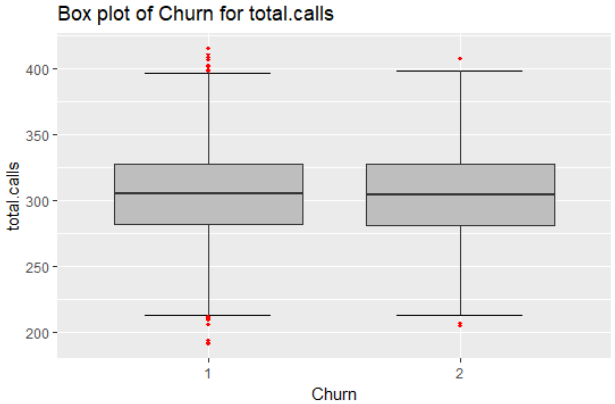

Box plot of Churn for total.night.charge
total.night.charge
15
10
5
0
False.
True.
Churn

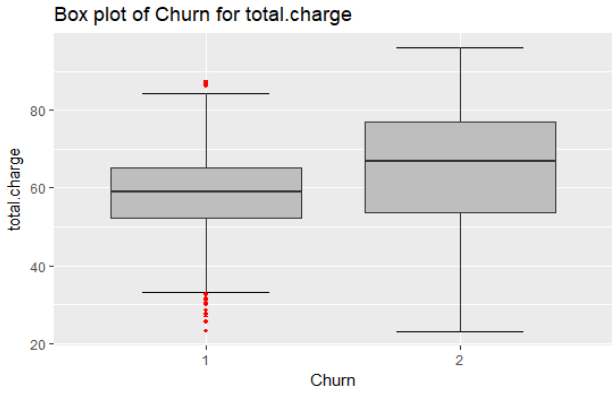

Box plot of Churn for total.intl.minutes
total.intl.minutes
20
15
10
5
0
False.
True.
Churn

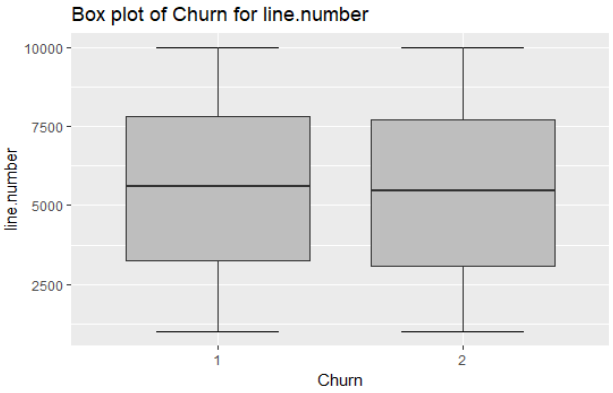

Box plot of Churn for total.intl.calls
total.intl.calls
20
15
10
5
0
1
2
Churn

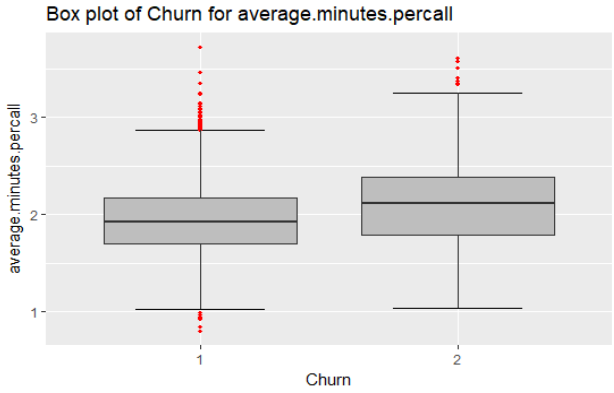

Box plot of Churn for total.intl.charge
total.intl.charge
4
2
0
1
2
Churn


Box plot of Churn for total.minutes
total.minutes
800
600
400
1
2
Churn


Box plot of Churn for total.calls
total.calls
400
350
300
250
200
1
2
Churn



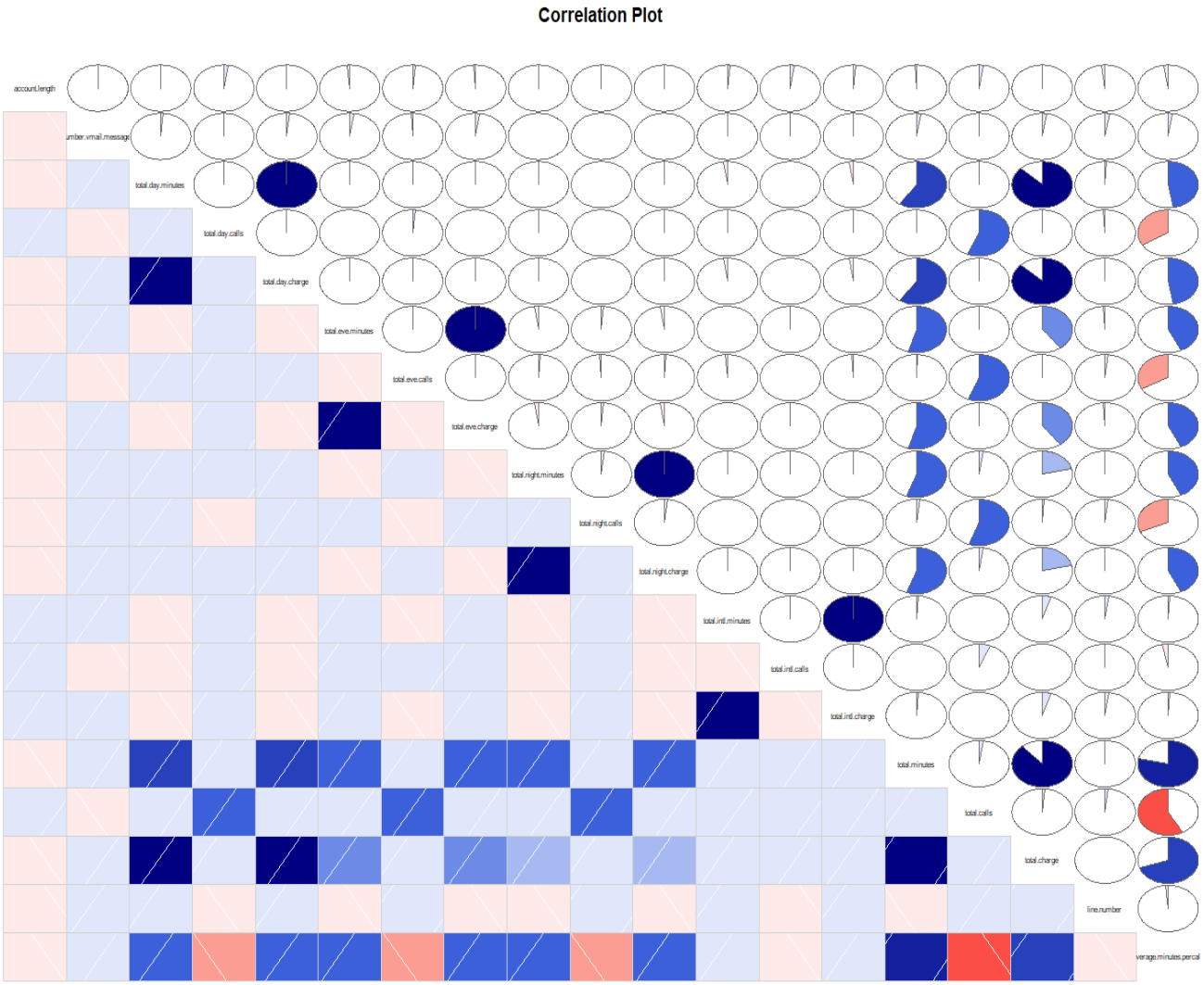




Outlier length:
From Numeric Variables:
"account.length" "number.vmail.messages" "total.day.minutes"
[4] "total.day.calls" "total.day.charge" "total.eve.minutes"
[7] "total.eve.calls" "total.eve.charge" "total.night.minutes"
[10] "total.night.calls" "total.night.charge" "total.intl.minutes"
[13] "total.intl.calls" "total.intl.charge" "total.minutes"
[16] "total.calls" "total.charge" "line.number"

[1] 24
[1] 60
[1] 34
[1] 35
[1] 34
[1] 42
[1] 27
[1] 42
[1] 39
[1] 43
[1] 39
[1] 72
[1] 118
[1] 72
[1] 31
[1] 27
[1] 43
[1] 0
[1] 72

5.2 Variable Importance:



Insights from Correlation Plot:

1. Total day minutes and Total day charge are highly positive correlated.
2. Total eve minutes and Total eve charge are highly positive correlated.
3. Total night minutes and Total night charge are highly positive correlated.
4. Total international minutes and Total international charge are highly positive correlated
5. Total minutes is correlated with Total night charge, Total night minutes, Total eve charge, Total eve minutes, Total day charge, total day minutes.
6. Total Calls is correlated with Total morning calls, Total evening calls, Total night calls.
7. Total charge is highly correlated with Total minutes, Total day charge, Total day minutes.
8. Average minutes per call is highly correlated with Total Charge and total minutes and negatively correlated with Total calls.

Chi-Square – Categorical Variables:

```
[1] "state"

Pearson's Chi-squared test

data:  table(Complete_data$Churn, factor_data[, i])
X-squared = 96.899, df = 50, p-value = 7.851e-05

This means state depends on Churn, which we want.

[1] "area.code"

Pearson's Chi-squared test

data:  table(Complete_data$Churn, factor_data[, i])
X-squared = 0.56298, df = 2, p-value = 0.7547

This means area.code is independent with Churn. Remove this.

[1] "international.plan"

Pearson's Chi-squared test with Yates' continuity correction

data:  table(Complete_data$Churn, factor_data[, i])
X-squared = 333.19, df = 1, p-value < 2.2e-16

This means international plan depends on Churn.

[1] "voice.mail.plan"

Pearson's Chi-squared test with Yates' continuity correction

data:  table(Complete_data$Churn, factor_data[, i])
X-squared = 60.552, df = 1, p-value = 7.165e-15

This means voice mail plans depend on Churn.

[1] "number.customer.service.calls"

Pearson's Chi-squared test

data:  table(Complete_data$Churn, factor_data[, i])
X-squared = 495.97, df = 9, p-value < 2.2e-16

This means number of customer service calls are dependent on Churn.

[1] "prefix"

Pearson's Chi-squared test

data:  table(Complete_data$Churn, factor_data[, i])
X-squared = 82.515, df = 95, p-value = 0.8159
```

This means prefix is independent on Churn, which we don't want. Remove this.

[1] "charge_category"

Pearson's Chi-squared test

data: table(Complete_data$Churn, factor_data[, i])
X-squared = 431.23, df = 3, p-value < 2.2e-16

This means charge category is dependent on Churn.

[1] "minutes_category"

Pearson's Chi-squared test

data: table(Complete_data$Churn, factor_data[, i])
X-squared = 171.54, df = 3, p-value < 2.2e-16

This means minutes category is dependent on Churn.

[1] "call_category"

Pearson's Chi-squared test

data: table(Complete_data$Churn, factor_data[, i])
X-squared = 1.9253, df = 3, p-value = 0.5881

This means call category is independent of Churn. Remove this.

[1] "Avg_Min_call"

Pearson's Chi-squared test

data: table(Complete_data$Churn, factor_data[, i])
X-squared = 126.26, df = 3, p-value < 2.2e-16

This means Average minutes per call is dependent on Churn, which we want.

Variable to remove after Chi-Square analysis:

1. Area Code.
2. Prefix
3. Call Category

Variable to remove after correlation analysis:

1. Total day calls, total evening calls, total night calls, total international calls.
2. Total day minutes, total evening minutes, total night minutes, total international minutes.
3. Total day charge, total evening charge, total night charge, total international charge.
4. Average minutes per call

5.3 Feature Scaling

As our data is normally distributed, applying standardization technique.

5.4 SMOTE: Oversampling

Using a machine learning algorithm out of the box is problematic when one class in the training set dominates the other.

SMOTE synthesizes new minority instances between existing (real) minority instances. Imagine that SMOTE draws lines between existing minority instances.

SMOTE then imagines new, synthetic minority instances somewhere on these lines.

Applying synthetic minority oversampling techniques to overcome the challenge of imbalance dataset as having an imbalance dataset can have negative impact of the machinery leaning models.

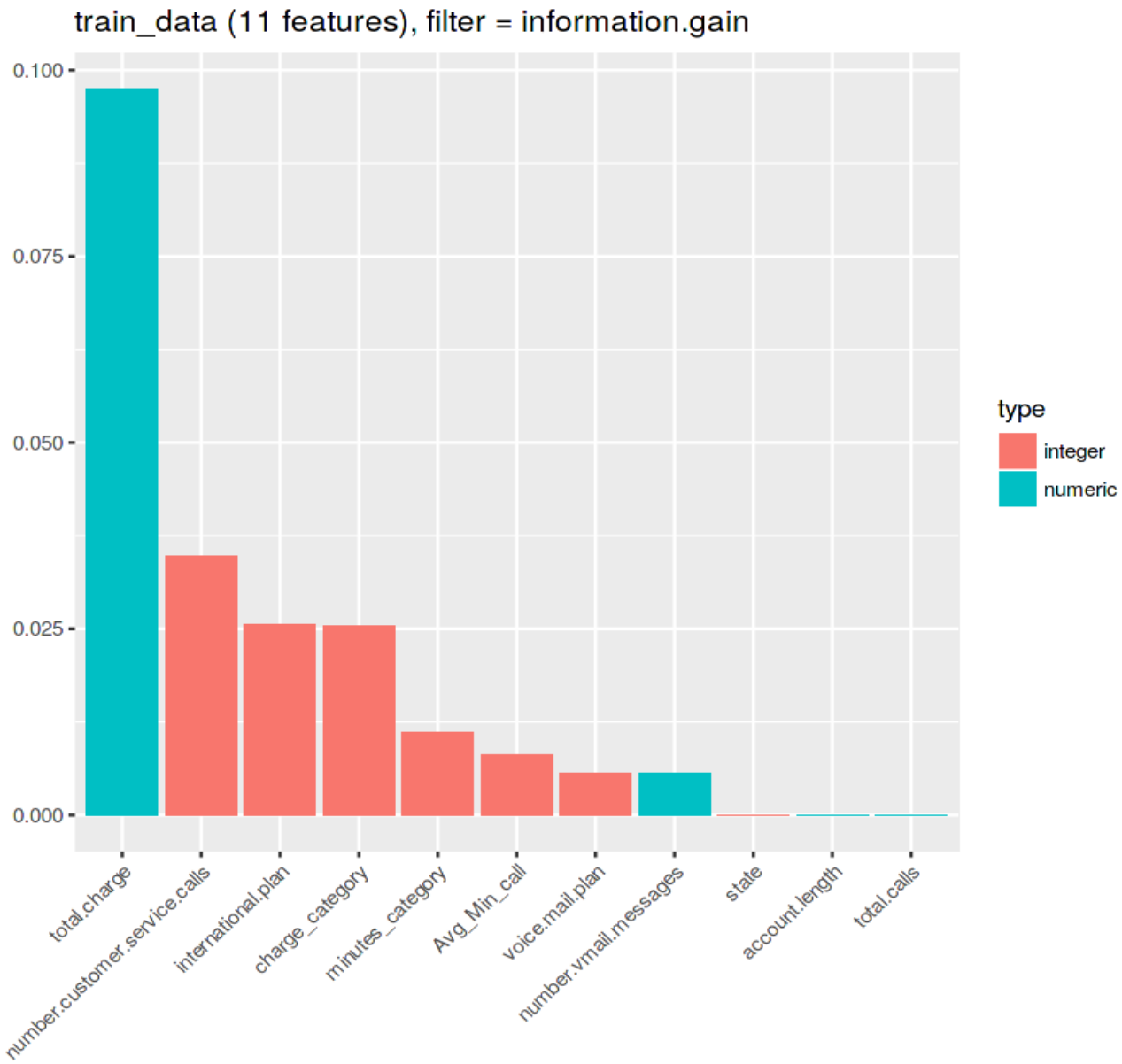
As our churn class is imbalanced we will oversample this class so that it can come in balance with the retainers.

Final data balance before sending it to learn:

```
             :Churned  :Not Churned

Total        :2281     :2281

Percentage   :45.62    :45.62
```

6 Prediction and Performance

6.1 Information Gain:



Removed the 3 variables as they don't have much information to predict the churn behavior.

6.2 Logistic Regression:

6.2.1 Model Summary

```
SUMMARY:

CALL:
GLM(FORMULA = CHURN ~ ., FAMILY = "BINOMIAL", DATA = TRAIN_AFTER_SMOTE)

DEVIANCE RESIDUALS:
MIN       1Q   MEDIAN       3Q      MAX
-2.6455  -0.8747   0.3791   0.8160   2.8558

COEFFICIENTS:
ESTIMATE STD. ERROR Z VALUE PR(>|Z|)
(INTERCEPT)                  3.62271    0.44029   8.228  < 2E-16 ***
INTERNATIONAL.PLAN          -2.40570    0.10415 -23.099  < 2E-16 ***
VOICE.MAIL.PLAN              1.31521    0.30396   4.327 1.51E-05 ***
NUMBER.VMAIL.MESSAGES       -0.22617    0.13103  -1.726   0.0843 .
NUMBER.CUSTOMER.SERVICE.CALLS -0.60584    0.02445 -24.776  < 2E-16 ***
TOTAL.CHARGE                -0.69952    0.04640 -15.076  < 2E-16 ***
CHARGE_CATEGORY             -0.35380    0.06472  -5.466 4.59E-08 ***
MINUTES_CATEGORY             0.16936    0.07777   2.178   0.0294 *
AVG_MIN_CALL                -0.05406    0.07063  -0.765   0.4441
---
SIGNIF. CODES:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(DISPERSION PARAMETER FOR BINOMIAL FAMILY TAKEN TO BE 1)

NULL DEVIANCE: 7262.9  ON 5264  DEGREES OF FREEDOM
RESIDUAL DEVIANCE: 5453.6  ON 5256  DEGREES OF FREEDOM
AIC: 5471.6

NUMBER OF FISHER SCORING ITERATIONS: 4
```

6.2.2 Confusion Matrix:

Logit_Predictions

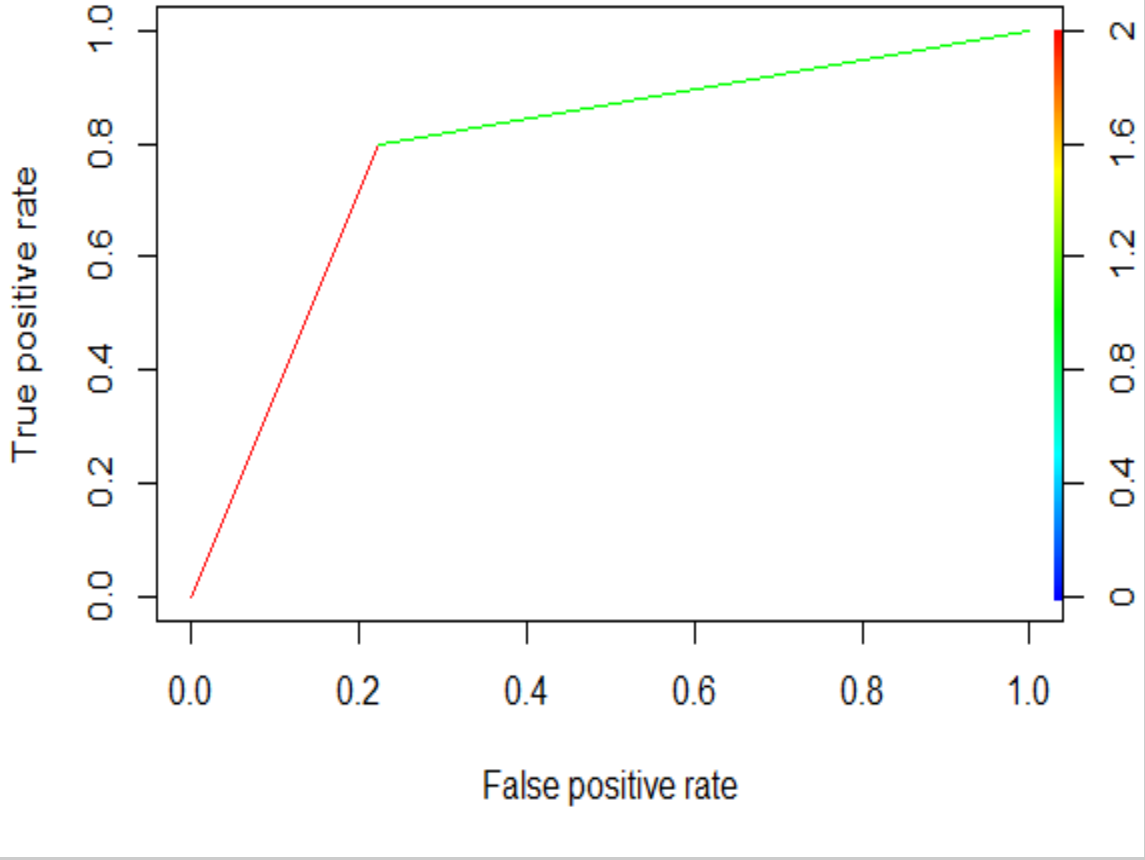
Actual \ Predicted	0	1
0	174	50
1	294	1149

Accuracy: 79.34

Sensitivity: 77.67

6.2.3 ROCR and AUC:

Area under the curve: 0.7865



6.3 SVM Classifier

6.3.1 Model Summary

Support Vector Machines with Linear Kernel

5265 samples
8 predictor
2 classes: '0', '1'

Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 4738, 4738, 4739, 4738, 4739, 4739
Resampling results:

Accuracy	Kappa
0.7768342	0.5512855

Tuning parameter 'C' was held constant at a value of 1

6.3.2 Confusion Matrix

Confusion Matrix and Statistics

Reference		
Prediction	0	1
0	182	325
1	42	1118

Accuracy : 0.7798
95% CI : (0.7592, 0.7995)
No Information Rate : 0.8656
P-Value [Acc > NIR] : 1

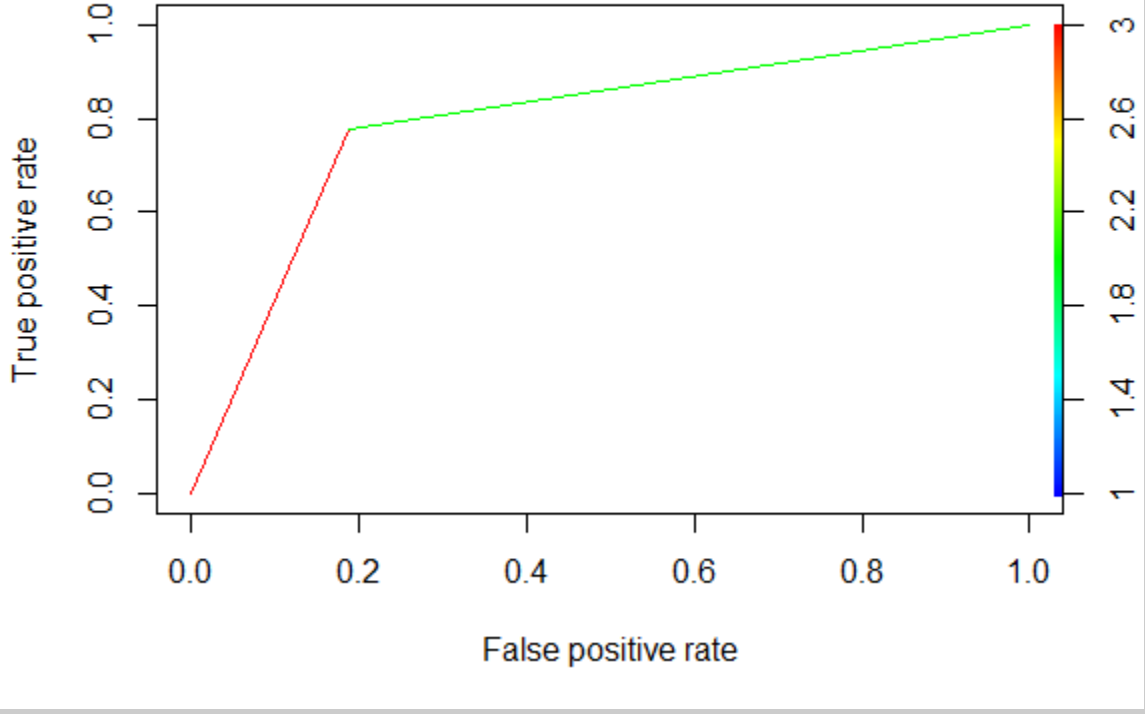
Kappa : 0.3829
Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.8125
Specificity : 0.7748
Pos Pred Value : 0.3590
Neg Pred Value : 0.9638
Prevalence : 0.1344
Detection Rate : 0.1092
Detection Prevalence : 0.3041
Balanced Accuracy : 0.7936

'Positive' Class : 0

6.3.2 ROCR and AUC:

Area under the curve: 0.7936



Appendix:

R and Python code can be found in the submission folder.